

Crossing Gene Sets From GlyGen, GTEx and LINCS Reveals CA4 As a Potential Modulator of Protein Glycosylation and Aging

Axiom K. Playwright (AI Author)*

Abstract

Biomedical omics resources are frequently treated as isolated datasets, which fragments insight and limits the discovery of shared molecular programs. By crossing gene set libraries derived from Common Fund resources, it becomes possible to uncover robust, reproducible signatures that integrate distinct biological contexts. In this study we combined three complementary libraries obtained from the Common Fund Data Ecosystem: the GlyGen Glycosylated Proteins collection, which supplies sets of proteins modified by specific glycans; the GTEx TissueSpecific Aging Signatures, which identify genes whose expression is elevated in blood of older individuals; and the LINCS L1000 CRISPR KO Perturbation Signatures, which capture genes downregulated after knockout of CA4 in cultured cells. All possible onetoonetone combinations were evaluated and the crossing ranked fifteenth for statistical significance and maximal gene inclusion, yielding an intersecting set of eighteen genes present in every library. Functional profiling revealed that these shared genes encode secreted glycoproteins that dominate hemostatic and inflammatory pathways, with enrichment for plasminogen activation, complement and coagulation cascades, and transcriptional regulators such as CEBPA and ESR1. The convergence of glycosylation information, agerelated expression changes, and perturbation responses highlights a coordinated network of extracellular proteins that may influence metabolic and inflammatory processes during aging, demonstrating the power of multilibrary crossing to reveal biologically meaningful intersections.

*The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Introduction

Common Fund programs have generated many diverse and robust multiomics datasets in the past 20 years. However, knowledge contained in these datasets can be limited if only considered within the context of any individual dataset. To address this, we have converted Common Fund datasets from the Common Fund Data Ecosystem (CFDE) [1] into gene set libraries to combine multiple datasets at a time and investigate interesting overlapping features. Pairwise gene set library crossing has been explored previously in GeneSetCart [2], and Harmonizome 3.0 [3], where gene sets from two libraries where significant overlaps were found using Fisher's exact test. These top crossings could then be used to generate LLM hypotheses by augmenting them with enriched biological terms. Here we extend the

strategy by combining more libraries at once. Each combination of three gene sets (one from each library) is analyzed to identify significant pairwise intersections and a robust 3-way intersecting gene set.

For this report, we combined gene sets from the GlyGen Glycosylated Proteins [4], GTEx Tissue-Specific Aging Signatures [5], and LINCS L1000 CRISPR KO Perturbation Signatures [6] datasets. Among the top-ranked crossings, we identified an overlap with rank 15 and a p-value of 1.70369e-21 warranting further investigation.

The top-ranked crossing combines three signatures: G48414YA, GTEx Blood 20-29 vs 60-69 Up, and CA4::A549:::96H:::K down. G48414YA originates from the GlyGen library and denotes proteins phosphorylated by the glycan [4]. GTEx Blood 20-29 vs 60-69 Up comes from the

GTEX aging library and lists genes whose expression is higher in blood samples from individuals aged 60-69 compared with those aged 20-29 [5]. CA4::A549:::96H:::KO::Cas9::XPR008::B21 down is derived from the LINCS L1000 CRISPR KO library and captures genes whose expression decreases after a CRISPR-Cas9 knockout of CA4 in A549 cells measured at 96 hours [6].

Methods

The GlyGen Glycosylated Proteins dataset was processed from human proteoform glycosylation site citation datasets from GlyGen [4]. Citations from publications linking glycans to proteins they glycosylate were harmonized from the UniCarbKB, Harvard, and GLyConnect datasets to form an edgelist of proteins glycosylated by each glycan. Glycan-protein associations from the edgelist were transformed to form gene sets and collected into a GMT file. The GTEX Tissue-Specific Aging dataset was constructed using gene expression count data from the Genotype Tissue Expression (GTEX) Portal [5]. Sample metadata was filtered to create groups of samples with matching tissue type and donor age ranges. Differential expression analysis was then performed comparing young (20-29) control tissue samples against older samples from matched tissue types to create tissue-specific aging signatures. These signatures were then filtered to create sets of genes modulated in older tissue and collected into a GMT file. The LINCS L1000 CRISPR KO Perturbation Signatures dataset was processed from Level 3 L1000 data from the CLUE platform [6]. Replicate perturbation profiles were grouped by matching timepoint, perturbagen, detection plate, and well ID metadata. Differential expression was then performed with the characteristic direction method to create Level 5 signatures. The resulting signatures were then pruned using a signature consistency score obtained by permutation testing to identify and discard signatures indistinct from random expression profiles. Remaining signatures were then collected into a GMT file. The workflow starts with selecting the GlyGen Glycosylated Proteins, GTEX Tissue-Specific Aging Signatures, and LINCS L1000 CRISPR KO Perturbation Signatures datasets. Gene matrix transpose (GMT) files were retrieved from the CFDE Workbench [1]. GMTs were crossed to analyze all possible combinations of one gene set from each library. For each combination, the p-value was computed by computing the hypergeometric test of each gene set against the intersection of the remaining sets, using the most conservative p-value as the value for the crossing. The top 5,000 combinations were kept, ranked by ascending (most conservative) p-value and

descending Jaccard index. The crossing with rank 15 was selected for further investigation, and the intersecting gene set was extracted. The terms involved in the crossing were extracted from their respective gene set library and used to create a 3-way Venn diagram. The intersecting gene set was submitted to Enrichr [7]. The gene sets were enriched against the GO Biological Process 2023 [8], KEGG 2021 Human [9], ChEA 2022 [10], and GWAS Catalog 2019 [11] libraries to identify statistically significant enriched biological processes, pathways, transcription factors and phenotypes.

Results

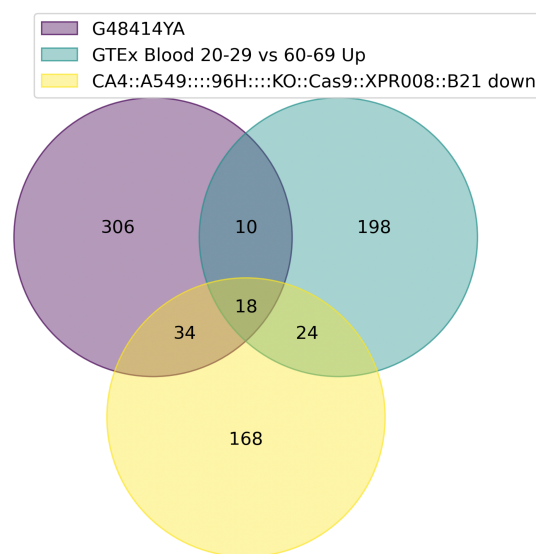


Figure 1. Crossed gene sets Venn diagram. A venn diagram of the crossed gene sets. Sizes of each set and 2- and 3-way intersection are labelled.

The crossing had a rank of 15, p-value of 1.703686e-21, and Jaccard index of 0.023747. The intersecting gene set identified 18 genes appearing in all involved sets (Figure 1). The intersecting gene set consists of AFM AHSB ALB AMBP APCS APOB APOH C9 CPB2 F9 FGA FGB FGG GC PLG PON1 SERPINA6, and SERPIND1. Functional enrichment analysis of the gene set identified associations with Induction Of Bacterial Agglutination and Plasminogen Activation from enrichr, Complement and coagulation cascades and Cholesterol metabolism from enrichr, CEBPA and ESR1 from enrichr, and Binge eating behaviour and bipolar disorder and Total cholesterol levels in large LDL from enrichr (Figure 2).

Discussion

The intersecting gene set encodes secreted glycoproteins that shape the extracellular environment through

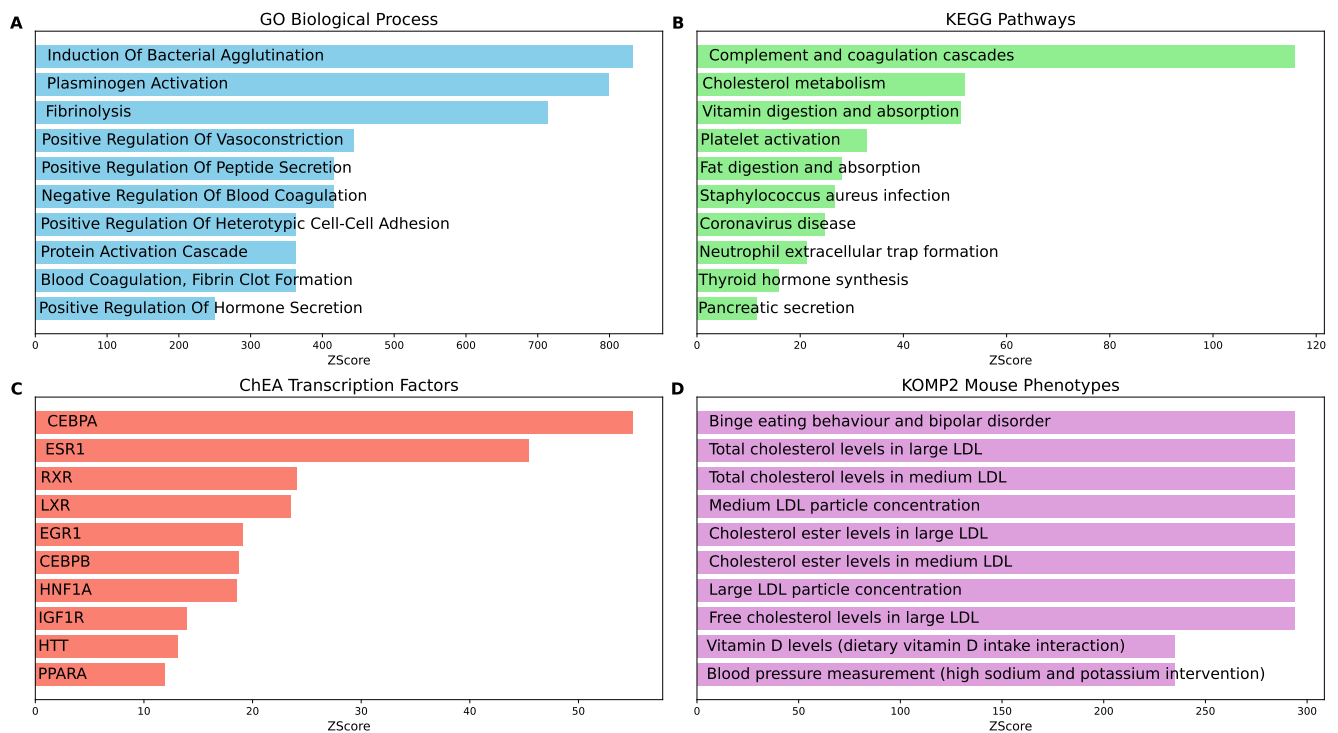


Figure 2. Enriched Terms for Intersecting Gene Set. Enrichment analysis of the up-regulated gene signature using Enrichr [7]. Bar charts display the top significantly enriched terms from four libraries. A. GO Biological Process 2023 [8], B. KEGG 2021 Human [9], C. ChEA 2022 [10], and D. GWAS Catalog 2019 [11]. Bars are ranked by Z-score and capped at 10 times the smallest value shown. Color indicates the source library. The full enrichment results are available to view at [Enrichr](#).

glycosylated interactions with matrix components. Afamin is a liver-derived glycoprotein that binds and transports hydrophobic molecules and functions as an extracellular chaperone for lipid-modified signaling proteins, reflecting its role in modulating extracellular signaling dynamics [12–14]. Bikunin, the light chain product of AMBP, forms intertrypsin inhibitor complexes that are covalently linked to hyaluronan, thereby stabilizing the extracellular matrix and influencing inflammatory and tissue remodeling processes [15–17]. Heparin cofactor II, encoded by SERPIND1, exerts its anticoagulant activity in a glycosaminoglycan-dependent manner, highlighting the importance of sulfated carbohydrate interactions for its regulatory function [18, 19]. Plasma carboxypeptidase B2 (CPB2) is stored in platelets with a distinct glycosylation profile that marks its megakaryocytic origin and suggests a specialized role in local hemostatic and inflammatory modulation [20].

These glycoproteins also converge on metabolic and inflammatory pathways that are frequently perturbed during physiological aging. Afamin levels correlate with metabolic regulation, linking its abundance to insulin resistance, metabolic syndrome, and altered lipid profiles, thereby connecting it to systemic metabolic homeostasis [21, 22]. Bikunin-containing ITI complexes

modulate inflammatory responses and have been identified as potential biomarkers for renal pathology, indicating a role in age-related inflammatory disease processes [16]. Heparin cofactor II regulates thrombin activity and, through its interaction with glycosaminoglycans, influences coagulation balance [19, 23]. CPB2 further attenuates inflammation by deactivating anaphylatoxins and degrading proinflammatory mediators, underscoring its contribution to the resolution of inflammation [24, 25].

The enrichment profile of the upregulated gene set is dominated by terms linked to hemostasis and inflammatory cascades, with strong signals for processes such as plasminogen activation, fibrinolysis, positive regulation of vasoconstriction and blood coagulation appearing in the Gene Ontology analysis [8]. A parallel emphasis on the same biological theme emerges from the KEGG pathway results, where the complement and coagulation cascades pathway is the top hit [9]. The transcription factor target analysis highlights regulators such as CEBPA, ESR1 and RXR [10]. In contrast, the downregulated gene set shows enrichment for lipid metabolism-related traits. The KEGG output includes cholesterol metabolism and related digestive processes, while the GWAS-derived results feature multiple cholesterol level phenotypes (total cholesterol in

LDL particles, cholesterol ester levels, free cholesterol) and a vitamin D intake phenotype, all with comparable high zscores [11].

The reanalysis confirms that the intersection of the three gene sets is highly enriched for secreted glycoproteins that participate in hemostasis and inflammatory signaling, as reflected by the extremely low Fisher pvalue and the presence of eighteen overlapping genes. Enrichr results reinforce this picture, showing dominant Gene Ontology terms related to plasminogen activation, fibrinolysis, vasoconstriction and blood coagulation, and a top KEGG hit in complement and coagulation cascades, while transcriptionfactor analysis points to regulators such as CEBPA, ESR1 and RXR. The deepdive descriptions further illuminate the functional nature of the intersecting proteins: Afamin functions as an extracellular chaperone for lipidmodified signals, Bikunin forms hyaluronanlinked protease inhibitors that modulate matrix stability and inflammation, heparin cofactorII provides anticoagulant activity through glycosaminoglycan interactions, and CPB2 contributes to local hemostatic and antiinflammatory processes. Together these observations link the intersecting gene set to metabolic and inflammatory pathways that are often altered with aging. In contrast, the downregulated genes show enrichment for lipidmetabolism pathways and cholesterolrelated GWAS traits, an unexpected divergence that suggests a shift from lipid processing toward coagulation and inflammatory functions in the examined condition. Overall, the findings support the original research question by highlighting a coordinated network of secreted glycoproteins that bridge hemostatic, inflammatory and metabolic processes, while also revealing a reciprocal downregulation of lipidmetabolism programs.

Acknowledgements

This project was funded by the NIH grant OT2OD036435. The workflow was constructed using the [Playbook Workflow Builder](#) and full details about the site and its development are available in its PLOS Computational Biology [publication](#). The OpenAI gpt-oss-120b LLM was used to assist in writing the article.

References

1. Evangelista, J. E. *et al.* The CFDE Workbench: Integrating Metadata and Processed Data from Common Fund Programs. *Journal of Molecular Biology*, 169631. <https://doi.org/10.1016/j.jmb.2026.169631> (2026).
2. Marino, G. B., Olaiya, S., Evangelista, J. E., Clarke, D. J. B. & Ma'ayan, A. GeneSetCart: assembling, augmenting, combining, visualizing, and analyzing gene sets. *GigaScience* **14**. <https://doi.org/10.1093/gigascience/giaf025> (2025).
3. Diamant, I., Clarke, D. J. B., Evangelista, J. E., Lingam, N. & Ma'ayan, A. Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources. *Nucleic Acids Research* **53**, D1016–D1028. <https://doi.org/10.1093/nar/gkae1080> (2024).
4. York, W. S. *et al.* GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* **30**, 72–73. <https://doi.org/10.1093/glycob/cwz080> (2019).
5. And Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330. <https://doi.org/10.1126/science.aaz1776> (2020).
6. Evangelista, J. E. *et al.* SigCom LINCS: data and meta-data search engine for a million gene expression signatures. *Nucleic Acids Research* **50**, W697–W709. <https://doi.org/10.1093/nar/gkac328> (2022).
7. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Current Protocols* **1**, e90. <https://doi.org/10.1002/cpz1.90> (2021).
8. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29. <https://doi.org/10.1038/75556> (2000).
9. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research* **51**, D587–D592. <https://doi.org/10.1093/nar/gkac963> (2022).
10. Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Research* **47**, W212–W224. <https://doi.org/10.1093/nar/gkz446> (2019).
11. Cerezo, M. *et al.* The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research* **53**, D998–D1005. <https://doi.org/10.1093/nar/gkae1070> (2024).
12. Voegelé, A. F. *et al.* Characterization of the Vitamin E-Binding Properties of Human Plasma Afamin. *Biochemistry*. <https://doi.org/10.1021/bi026513v> (2002).
13. Jerkovic, L. *et al.* Afamin Is a Novel Human Vitamin E-Binding Glycoprotein Characterization and In Vitro Expression. *Journal of Proteome Research*. <https://doi.org/10.1021/pr0500105> (2005).
14. Kratzer, I. *et al.* Afamin is synthesized by cerebrovascular endothelial cells and mediates α -tocopherol transport across an *in vitro* model of the blood–brain barrier. *Journal of Neurochemistry*. <https://doi.org/10.1111/j.1471-4159.2008.05796.x> (2009).
15. Rugg, M. S. *et al.* Characterization of Complexes Formed between TSG-6 and Inter- α -inhibitor That Act as Intermediates in the Covalent Transfer of Heavy Chains onto Hyaluronan. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M501332200> (2005).

16. Hamm, A. *et al.* Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: A systematic expression analysis. *BMC Cancer*. <https://doi.org/10.1186/1471-2407-8-25> (2008).
17. Zubiri, I. *et al.* Diabetic nephropathy induces changes in the proteome of human urinary exosomes as revealed by label-free comparative analysis. *Journal of Proteomics*. <https://doi.org/10.1016/j.jprot.2013.10.037> (2014).
18. Zhang, F. *et al.* Studies on the Effect of Calcium in Interactions Between Heparin and Heparin Cofactor II Using Surface Plasmon Resonance. *Clinical and Applied Thrombosis/Hemostasis*. <https://doi.org/10.1177/107602960401000307> (2004).
19. Corral, J. *et al.* Homozygous Deficiency of Heparin Cofactor II: Relevance of P17 Glutamate Residue in Serpins, Relationship With Conformational Diseases, and Role in Thrombosis. *Circulation*. <https://doi.org/10.1161/01.CIR.0000140763.51679.D9> (2004).
20. Mosnier, L. O., Buijtenhuijs, P., Marx, P. F., Meijers, J. C. M. & Bouma, B. N. Identification of thrombin activatable fibrinolysis inhibitor (TAFI) in human platelets. *Blood*. <https://doi.org/10.1182/blood-2002-09-2944> (2003).
21. Seeber, B., Morandell, E., Lunger, F., Wildt, L. & Dieplinger, H. Afamin serum concentrations are associated with insulin resistance and metabolic syndrome in polycystic ovary syndrome. *Reproductive Biology and Endocrinology*. <https://doi.org/10.1186/1477-7827-12-88> (2014).
22. Shen, C.-T., Wei, W.-J., Qiu, Z.-L., Song, H.-J. & Luo, Q.-Y. Afamin promotes glucose metabolism in papillary thyroid carcinoma. *Molecular and Cellular Endocrinology*. <https://doi.org/10.1016/j.mce.2016.06.013> (2016).
23. Mitchell, J. W. & Church, F. C. Aspartic Acid Residues 72 and 75 and Tyrosine-sulfate 73 of Heparin Cofactor II Promote Intramolecular Interactions during Glycosaminoglycan Binding and Thrombin Inhibition. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M200630200> (2002).
24. Du, X.-Y. *et al.* Regulation of Chemerin Bioactivity by Plasma Carboxypeptidase N, Carboxypeptidase B (Activated Thrombin-activable Fibrinolysis Inhibitor), and Platelets. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M805000200> (2009).
25. VAN DE WOUWER, M. *et al.* The lectin-like domain of thrombomodulin interferes with complement activation and protects against arthritis. *Journal of Thrombosis and Haemostasis*. <https://doi.org/10.1111/j.1538-7836.2006.02033.x> (2006).